Collibra

Collibra Data Intelligence Cloud

Cloud Infrastructure

Version 2022.05

Collibra

Collibra Data Intelligence Cloud - Cloud Infrastructure

Release date: May 8, 2022

Revision date: Wed May 04, 2022

You can find the most up-to-date technical documentation on our Documentation Center at

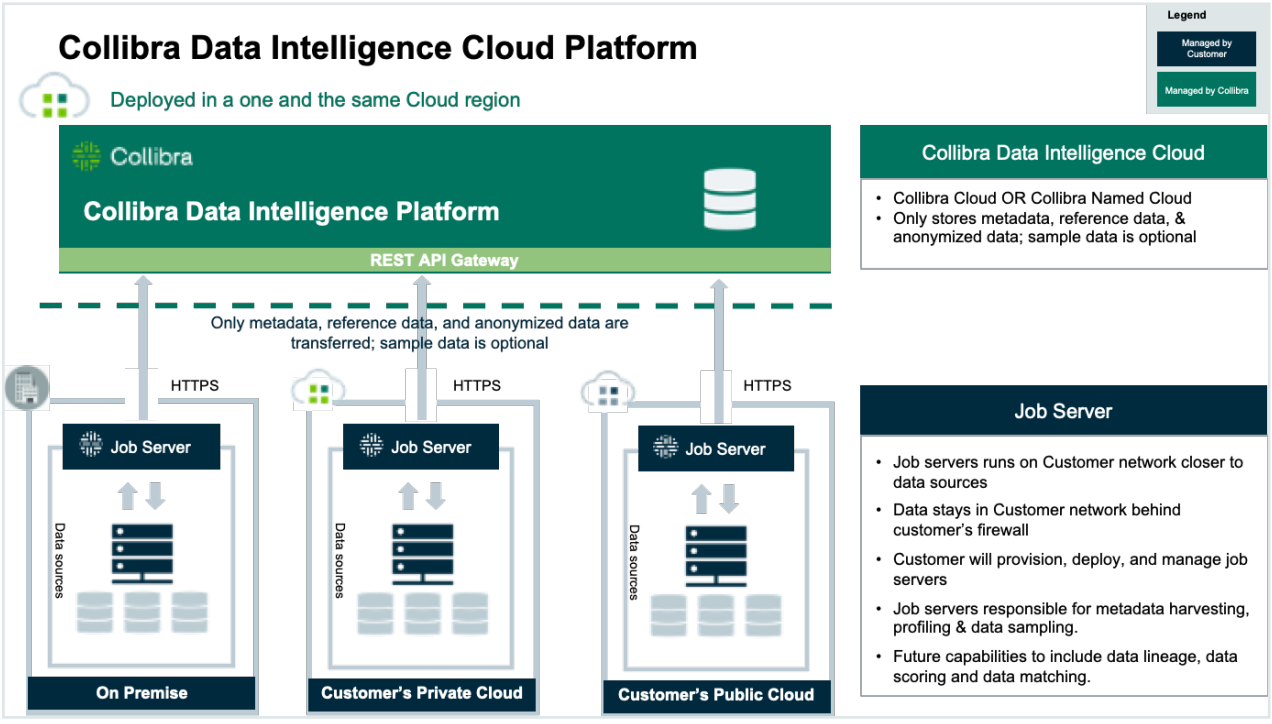
https://productresources.collibra.com/docs/collibra/latest/Content/Cloud/to_cloud-infrastructure.htm

Contents

Collibra Data Intelligence Cloud

The Collibra Data Intelligence Cloud is delivered as a Software as a Service offering. This guide describes the Collibra Cloud Infrastructure processes including network security, backup and recovery processes, monitoring and security.

The following diagram shows how you connect to Collibra Data Intelligence Cloud using the Jobserver server installed on the cloud premises or cloud. Collibra Data Intelligence Cloud is where you perform all of your work. The Jobserver connects to your data sources to perform actions like data profiling and then provides that information back to Collibra Data Intelligence Cloud.



Collibra Cloud Infrastructure overview

This section provides an overview of the infrastructure and architecture of a Collibra Cloud.

Cloud regions and availability zones

Collibra has a cloud presence in many places around the world. Cloud centers are located so that each customer's data can be kept in the appropriate regulatory region. The on-boarding process includes selection of an appropriate regulatory region based on each unique customer's requirements. Under no circumstances does a customer's data ever leave the regulatory region agreed upon. Data may be moved around for resilience or performance reasons within a regulatory region.

You can find an overview of regions in which Collibra operates in the next section.

Customer separation

Collibra uses a variety of technical means to keep each customer's data separate from every other customer's data. Separation is ensured during storage, internal transit, and processing at all times.

Collibra cloud hosting regions

The following table contains an overview of the hosting regions in which Collibra operates.

Cloud provider	Region name	Location
AWS	us-east-1	North Virginia (USA)
AWS	us-west-1	North California (USA)
AWS	eu-west-1	Ireland (Europe)
AWS	eu-west-2	London (Europe)
AWS	eu-central-1	Frankfurt (Europe)
AWS	ca-central-1	Canada Central (Montreal area of Quebec)
AWS	ap-southeast-2	Sydney, Australia
AWS	us-gov-west-1	US GovCloud West (FedRAMP)
AWS	ap-southeast-1	Singapore
GCP	us-east1	Moncks Corner, South Carolina, USA
GCP	us-east-4	Ashburn, Northern Virginia, USA
GCP	us-central1	Council Bluffs, Iowa, USA
GCP	northamerica-northeast1	Montreal, Quebec, Canada
GCP	europe-west1	St. Ghislain, Belgium
GCP	europe-west2	London, England, UK
GCP	australia-southeast1	Sydney, Australia
GCP	asia-southeast1	Jurong West, Singapore
Azure	Germany West Central	Frankfurt

Cloud provider	Region name	Location
Azure	US East 2	Virginia

Data privacy and security

Data privacy and security are an essential aspect of Collibra Data Intelligence Cloud.

Collibra is constantly improving its security procedures in order to guarantee the best possible security standards.

Cloud infrastructure

Collibra works closely with various Infrastructure-as-a-Service cloud providers to provide a flexible and secure environment. For Collibra's cloud offering, these services include:

- Data center
- Server hardware
- Network infrastructure
- Cloud provider specific services

Physical access to the servers is subject to the privacy statements of cloud providers.

Encryption

All data is encrypted in transit between the Jobserver (installed at client) and the Collibra Data Intelligence Cloud. It uses mutual authentication via certificates over TLS 1.2. When data is stored in Collibra, it is always encrypted at rest using AES 256.

Key management

Keys are managed by Collibra using the key management systems native to the cloud provider. The KMS used is FIPS 140-2 compliant 140-2, uses AES 256-bit encryption and the keys are rotated at standard intervals.

Collibra employees

Collibra is aware that, when it comes down to security, people are often the weakest link.

- Every Collibra employee signs a non-disclosure agreement (NDA) contract concerning all company and customer data.
- Before new employees are hired, background checks are done, to the best of our abilities.
- All desktops and laptops used by Collibra employees are encrypted using a unique (randomly generated) key.

- We have several roles at Collibra, and only qualified people will have access to the customer virtual machine.
- An employee termination process is in place in order to remove all access to Collibra internal services and data.

Password management

All customer data is protected by highly secure (random) passwords. Different passwords are used for:

- The virtual machine running the application.
- The database
- The administrator access to the application
- The backup encryption

Passwords are kept in a highly secure digital vault, AES-256 algorithms with encryption keys of at least 80 bit. This vault allows us to make specific passwords available to specific people.

Database, virtual machine and backup passwords are only accessible by senior management and infrastructure support employees.

Application administrator passwords are only accessible by senior management and application support employees.

Additionally, temporary access to the necessary passwords is provided to development or pre-sales employees to solve specific problems in case of an emergency.

Employee training

Collibra employee trainings are organized internally during the last quarter of every year. This training covers the following topics:

- Password vault usage
- Data Privacy and Data Security awareness and best practices
- Securing your desktop/laptop
- Cloud infrastructure

New employees are given this training within the first three months of their employment.

Chapter 4

Access and authentication

This section provides more information on access and authentication to Collibra Data Intelligence Cloud environments by customers and the Collibra Cloud Infrastructure team.

Customer access to Collibra cloud

You can only access Collibra Data Intelligence Cloud by going to https://<customer-name>.collibra.com.

> Note It's impossible to access the actual servers that contain the data. This also means that you cannot install other services and applications.

On request, it is possible to limit inbound access to a specific IP-range (IP Whitelisting). To enforce a limit, you have to create a ticket on the Collibra support portal.

Whitelisting IP addresses

By default, all source IP addresses on the Internet can reach your environment. Whitelisting is a cybersecurity enhancement that prevents unexpected sources from reaching your Collibra environment.

When whitelisting is enabled for your environments, all source IP addresses are blocked, except for the source IP addresses that you provided in your support ticket.

To avoid interruptions and access issues to your environments, ensure that you provide a comprehensive and validated list of your internal egress IP addresses. These egress IP addresses could include your corporate offices, VPN, integrations and partners; generally any client that should be able to access your environment.

We strongly recommend that you work with your internal departments to identify who needs access, and then apply the whitelisting on your non-production environments. This allows you to validate the list of source IP addresses while mitigating access interruptions to mission-critical or production environments.

IP address format

In your support ticket, you must provide a comma-separated list of all IP addresses that you want whitelisted, in Classless Inter-Domain Routing (CIDR) notation. The IP addresses should be routable.

Addresses in private spaces, such as 10.0.0.0 or 192.168.0.0, are not valid and will be ignored.

Collibra Cloud Infrastructure team access

The Collibra Cloud Infrastructure team can sign in to the cloud servers to perform necessary management and maintenance tasks. Collibra Security does not allow direct management access to the actual servers containing the data, and must be done through a dedicated management server and then through a bastion host.

Collibra has an Identity and Access Management (IAM) system in place to provide separation of rights. Only dedicated and trained personnel have access to the entire system. Each user has its own unique user ID to track responsibility of changes to the system. Access is handled through private/public key authentication secured by a FIPS-140 second factor.

Chapter 5

Monitoring a Collibra cloud environment

Collibra uses multiple monitoring services to provide a near-real-time visibility into the internal cloud infrastructure.

Monitoring alerts

Every customer cloud environment is monitored on a minute-by-minute basis from multiple external locations around the world. This monitoring checks the availability of the service and sends alerts to the Collibra Cloud Infrastructure team in case of an outage. An incident management process is then initiated to resolve the problem as soon as possible.

The Cloud Infrastructure team also uses a continuous monitoring solution internally to manage performance and capacity for every customer.

Monitoring uptime

A system run by Collibra fully monitors and reports on uptime. This system is able to make a difference between maintenance mode and other downtime to measure the uptime with higher accuracy.

Detailed reporting on uptime is available on a monthly basis at customer request .

Performance monitoring

Performance of the cloud resources is monitored using a specialized service.

This includes:

- Memory usage
- CPU utilization
- Disk usage
- Network usage
- (Average) Throughput (requests per minute)
- (Average) Response times

When performance problems occur for any of these resources, alerts are sent to the Collibra Cloud Infrastructure team so that prompt action can be taken.

Security monitoring

In order to track changes to our cloud servers, the following actions are logged:

- SSH authentications
- User commands
- Changing user privileges
- Sign-in attempts
- Other indicators specific to our internal platform

Host Intrusion Detection System

A Host Intrusion Detection System (HIDS) monitors network traffic for suspicious activity and alerts the system or network administrator. In some cases, the HIDS may also respond to anomalous or malicious traffic by taking action, such as blocking the user or source IP address from accessing the network.

The tools, as implemented by Collibra, detects the following intrusion possibilities:

- File integrity (system and application files)
- Sign-in attempts
- Portscanning
- Brute force attacks
- Rootkit detection

All alerts are collected in a central place, and alerts with high priority are reviewed by both Production Engineering and Security Operations.

Host Intrusion Prevention System

A Host Intrusion Prevention System (IPS) is a network security/threat prevention technology that examines network traffic flows to detect and prevent vulnerability exploits.

Collibra has an active response system in place that does not only send out alerts (HIDS), but also takes action (HIPS). The HIPS can take the following actions when needed:

- Deny IP addresses to access the system (firewall, hosts.deny).
- Disable the user accounts under attack.
- Drop packets.

Antivirus and malware

Collibra is always checking for the latest virus and malware (including rootkits) vulnerabilities. Our automated process will download the newest threats and search the system for possible virus and malware presence.

If a threat is detected, the customer is notified.

Chapter 6

Backup and recovery

To avoid data loss, Collibra implements a full and secure backup. Here is an overview of our backup and recovery processes.

Repository backups

Collibra backup system

The repository of a Collibra Data Intelligence Cloud environment is backed up every 15 minutes. With this backup policy, Collibra can restore your environment to 15 minutes accurate, limited to 30 days back in time. Each night, depending on the time zone of the server, Collibra takes a snapshot by freezing the volume. This assures continuous backups on a scheduled basis. With these backups, Collibra can restore your system or data as required throughout the subscription term.

At the end of the subscription term, the final backup is retained and available to the customer for 30 days after subscription termination.

Replication of backups

For Disaster Recovery purposes, Collibra ensures that your data is replicated across multiple data centers located in the same regulatory region. These backups are encrypted at rest using AES-256 encryption.

Recovery

The procedure to recover an encrypted backup depends on the event that triggers the need for recovery.

- Data loss / corruption: Files are restored to their last good state.
- Database corruption: The database dump needs to be analyzed to find the time stamp of corruption. After this investigation, Collibra can restore the last known snapshot.
- Server problems: Collibra analyzes the server problems. If the problem is not found in a reasonable time, then a full restore will be scheduled with the customer. The application and the data can be restored within eight business hours.
- Data center crash and loss of data: In the worst case, Collibra can restore an off-site backup in a different data center.

> Note The detailed internal procedures on how we perform a recovery, are not made publicly available.

Adding a Jobserver to Collibra Data Intelligence Cloud

When you want to ingest data, you need the Jobserver service in your environment, whether it's an on-premises environment or a cloud environment. From a network point of view, we recommend in both situations to install the Jobserver service as close to the data source as possible for optimal ingestion performance. This often means that you will have to install an on-premises Jobserver service

> Note If you use the Jobserver from your Collibra Data Intelligence Cloud environment, we recommend that you only use it for the ingestion of Excel or CSV files.
> Do not use this Jobserver to ingest databases, your environment can become unresponsive. To ingest databases, make sure that your Jobserver meets the system requirements.

This section describes how to install and add the Jobserver service to your Collibra Data Intelligence Cloud environment.

> Tip
> When you install an on-premises Jobserver for use in a Collibra Data Intelligence Cloud environment, you also have to install Collibra Console, to manage and configure this Jobserver. You can install both services on the same server.
>
> Additionally, you can install the Monitoring service to monitor the Jobserver service.
>
> You can find the version of your Collibra Data Intelligence Cloud environment at the bottom of the sign-in window, for example 2022.05.0. Always use the latest available on-premises installer to install the Jobserver.

Supported operating systems

Note
- Only 64-bit operating systems are supported.
- Linux operating systems are recommended over Windows operating systems.
- Windows Administrator rights with full rights on the intended installation drive/directories are mandatory.

Linux operating systems

- Red Hat Enterprise Linux/CentOS 6.x
- Red Hat Enterprise Linux/CentOS 7.x
- Debian 9
- Ubuntu 16.x
- Ubuntu 18.x
- Suse 12

Note
- You have to set the locale to *en_US.UTF-8* on all Linux systems.
- Root permissions are not mandatory but preferred.
  If you install the Jobserver without root permissions, see the services section.

Microsoft Windows operating systems

- Windows Server 2012 R2
- Windows Server 2016
- Windows Server 2019

Jobserver requirements

The following requirements are for an on-premises Jobserver that will be used with a Collibra Data Intelligence Cloud environment.

Data ingestion

- 64 GB RAM
- 500 GB free disk space
- Hard disk type: SSD
- Number of CPUs: 16

Tableau ingestion

- 6 GB RAM
- 35-50 GB free disk space
- Hard disk type: SSD
- Number of CPUs: 4

Collibra Console requirements

- 1 GB RAM
- 1 GB free disk space for the installation.
- 1 GB free disk space for the Collibra Console database.

As this Collibra Console is only used to manage the on-premises Jobserver, you don't need extra disk space for backups.

Check connectivity between an on-premises Jobserver and Collibra Data Intelligence Cloud

Before you install your on-premises Jobserver, we highly recommend to do some basic connectivity tests between the node on which you are going to install the Jobserver service and your Collibra Data Intelligence Cloud environment. If the node cannot reach your Collibra Data Intelligence Cloud environment, then first fix the connectivity before you start the Jobserver installation.

1. Open the connection and port from your on-premises Jobserver to your Collibra Data Intelligence Cloud environment.
2. Whitelist the DNS name of your Collibra Data Intelligence Cloud environment.
3. Check if the Jobserver node can reach your Collibra Data Intelligence Cloud environment with the following command:

   ```
   curl -x "" -i https://<your-environment>.collibra.com/reversehttp-poll/1
   ```

   You should receive an HTTP 401 response. If you don't receive an HTTP 401 response, fix the connectivity before proceeding with the configuration.

Tip If there is a proxy server between your on-premises Jobserver and your Collibra Data Intelligence Cloud environment, then add the proxy address to the command:

```
curl -x http(s)://proxy:port -i https://<your-environment>.collibra.com/reversehttp-poll/1
```

What's next?

When you receive an HTTP 401 response, you can start the installation of the Jobserver.

Install the Jobserver and Collibra Console on a single node

This section describes how to install the Jobserver and Collibra Console on a single node. There is no need to install both components on different servers.

Prerequisites

- You have downloaded the latest Jobserver-only installer for your operating system. See also the compatibility list to know which installer you have to download.

> Note You can still use the full on-premises installer but that contains all on-premises services.

- We recommend to use a static IP address for the node.
- You have a connection between the on-premises Jobserver and Collibra Data Intelligence Cloud.

Steps

Linux

1. Run the installer:
   - Linux as user with sudo rights: `sudo ./dgc-linux-jobserver-only-2022.5.0-[to be defined].sh`
     Linux as root user: `./dgc-linux-jobserver-only-2022.5.0-[to be defined].sh`
   - Linux as standard user: `./dgc-linux-jobserver-only-2022.5.0-[to be defined].sh`
2. Follow the command-line wizard. If you don't enter a value, the value between brackets or the value in capital is used.

> Note The Jobserver encryption key in the wizard is a passphrase that is used to generate the actual encryption key.

```
Verifying archive integrity...  100%   All good.
Uncompressing DGC Installer  100%
Specify the installation directory [/opt/collibra]:

Please specify the data directory [/opt/collibra_data]:

Do you want to install the Collibra Jobserver component?
[Y/n]

Do you want to install the Collibra Management Console com-
ponent? [Y/n]

Are you sure these are the components you want to install?
[Management Console, Jobserver] [Y/n]
```

```
Specify the Jobserver port [4404]:

Specify the Jobserver database port [4414]:

Specify the Jobserver monitoring port [4424]:

Specify the Jobserver Spark monitoring port [4434]:

Specify the Agent port [4401]:

Specify the Agent address [localhost]:
Note: with a loopback address (localhost, 127.0.0.1, et
al.) you will not be able to use a multi node setup

Specify the Management Console context path []:

Specify the Management Console port [4402]:

Specify the Management Console database port [4420]:

2021-02-15 04:51:34.302 - SUCCESS - Create user and group
2021-02-15 04:51:34.310 - SUCCESS - Check umask settings

...

2021-02-15 04:51:55.387 - SUCCESS - Start Console
2021-02-15 04:51:55.387 - COMPLETED - Installation finished
in 21419ms.
```

Windows

> Note Anti-virus and/or security software may block the installation on Windows. Make sure that these allow the installation of software and services.
> For more information, see also the Collibra University course.

1. Run the installer: Windows Server: double-click **setup.bat**The path of the installer file cannot contain spaces.
   If you run the installation without Administrator rights, an error is shown.
2. In the wizard introduction, click **Next**.
3. Enter the **Installation directory** and click **Next**.
4. Enter the **Data directory** and click **Next**.
5. Select **Management Console** and **Jobserver** and click **Next**.

6. Enter the Jobserver settings and click **Next**.

Setting	Description
Jobserver port	The TCP port to access the Jobserver service. The default port is *4404*.
Jobserver database port	The TCP port to access the Jobserver database. The default port is *4414*.
Jobserver monitoring port	The TCP port that is used by the monitoring service to monitor the Jobserver service. The default port is *4424*.
Jobserver Spark monitoring port	The TCP port that is used by the monitoring service to monitor the Jobserver Spark service. The default port is *4434*.

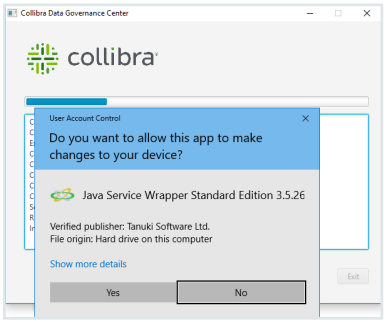
7. Enter the Console settings.

Setting	Description
Console port	The TCP port to access your Collibra Console via your web browser. The default port is *4402*.
Console database port	The TCP port to access the Collibra Console database. This is the database where the data and configuration of Collibra Console is stored. The default port is *4420*.

Setting	Description
Console context path	The path that is added to the base URL to reach Collibra Console. For example, if your base URL is `https://dgc.yourcompany.com:4402/` and your context path is `console-acceptance`, then your path to reach Collibra Console is `https://dgc.yourcompany.com:4402/console-acceptance`. See also Set the context path of Collibra Console.

If you run multiple Collibra Console instances on one node, this port must be unique for each instance.

8. Click **Install**.
   » The installation of the components starts.
9. On Windows, you may see User Account Control warnings requesting to make changes to your device.



   Click **Yes** for each of the requests, if you click **No**, the installation will fail.
10. Click **Exit**.
    » Collibra is installed on your system.

What's next?

> Tip If you don't install both services on the same node, then you have to add the Jobserver node to the on-premises Collibra Console.

Configure the connection between the Jobserver and your Collibra Data Intelligence Cloud environment by creating a connection from the Jobserver to the DGC service. In this configuration, the on-premises Jobserver will poll the DGC service for ingestion tasks.

> Note You can also add the Jobserver service to the DGC service. In this configuration, the DGC service from your cloud environment will send ingestion tasks to the on-premises Jobserver. For security reasons, this configuration may not be allowed by your organization's network infrastructure department

Connection from an on-premises Jobserver to a Collibra Data Intelligence Cloud

In default installations, a Jobserver is installed on-premises and a Collibra Data Intelligence Cloud sends ingestion and profiling jobs to it. However, we highly recommend to reverse this communication, so that the on-premises Jobserver polls for jobs. This is often required for security reasons.

> Note If you want a Collibra Data Intelligence Cloud to send jobs to your on-premises Jobserver, contact your security officer and network administrator.

In this section, you get more information on how to set up the communication from an on-premises Jobserver to a Collibra Data Intelligence Cloud.

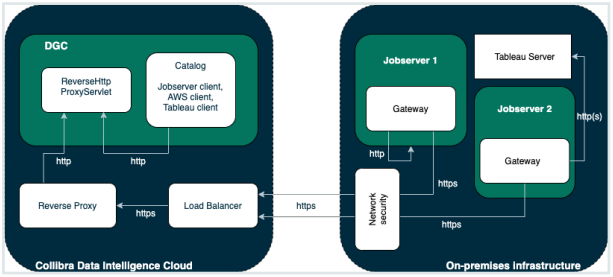
On-premises Jobserver to Collibra Data Intelligence Cloud communication

By default, the Data Governance Center service sends ingestion and profiling jobs to the Jobserver. This means that if you are using a Collibra Data Intelligence Cloud environment with an on-premises Jobserver, there is an inbound connection to the customer network, which is often not possible for security reasons. To allow such connections, you can use a reverse proxy server.

But, instead of using a reverse proxy, you also have to possibility to reverse this communication, where the on-premises Jobserver initiates the communication to Collibra Data Intelligence Cloud.

Communication overview

The following schema shows the communication paths from an on-premises Jobserver to Collibra Data Intelligence Cloud or to an on-premises Tableau server:



In case of an on-premises Tableau server, it is possible that inbound connections to it are not allowed. To establish a communication between your Collibra Data Intelligence Cloud and Tableau server, you will need a Jobserver that is dedicated to ingest from the Tableau server. The configuration of the communication from the Jobserver to the Tableau server is similar to the one from a Jobserver to a Collibra Data Intelligence Cloud environment.

Each Jobserver has to be a dedicated Jobserver, you cannot use a Jobserver to ingest from both Tableau server and S3 or JDBC data sources.

Components

To enable communication from an on-premises Jobserver to Collibra Data Intelligence Cloud, there are two new components:

New component	Description
Reverse HTTP proxy servlet	The reverse HTTP proxy is part of the DGC service. It acts as a server for all other Collibra services, whether they are installed on-premises or together with the DGC service in the cloud.

New component	Description
Gateway	The gateway is part of the Jobserver service. It polls the DGC service's reverse proxy to fetch tasks and send them to the Jobserver.

You only need the gateway to communicate with an on-premises Tableau server.

Configure the Jobserver to Collibra Data Intelligence Cloud communication

By default, Collibra Data Intelligence Cloud sends jobs to a Jobserver but you also have the possibility to have the Jobserver poll Collibra Data Intelligence Cloud for jobs.

If you want to have the Jobserver poll Collibra Data Intelligence Cloud for jobs, you have to configure both the Data Governance Center service and the Jobserver service.

Prerequisites

- You have Collibra Data Intelligence Cloud 2020.10 or newer.
- You have created a keystore in the PKCS#12 format on the node that hosts the Jobserver service.

Steps

Configure the Jobserver service

> Warning
> Only apply changes to these settings if you are really experienced with JVM parameters. Changing parameters may cause serious performance issues.
>
> Restart the service after editing the JVM parameters.

Execute the following steps in the Collibra Console instance that manages your Jobserver:

1. Open Collibra Console with a user profile that has the **SUPER** role.
   » Collibra Console opens with the **Infrastructure** page.

   > Tip The default address to access Collibra Console is *<server hostname>:4402*, but you may have set another port during the installation of Collibra Console. Keep in mind that a firewall of your operating system can block the access to Collibra Console.

2. In the tab pane, click the Jobserver service of a Collibra environment.
3. Click **Infrastructure Configuration**.
4. Click **JVM configuration**.
5. Click **Edit configuration**.
6. Add the following JVM settings:

Setting	Description
reversehttp.gateway	The setting to enable the Jobserver's gateway. To enable the communication from the on-premises Jobserver to the Collibra Data Intelligence Cloud environment., this value must be *true*. Example `-Dreversehttp.gateway=true`
proxy.url	The URL of the Collibra environment, followed by *reversehttp-poll/<gateway-id>*. This "gateway-id" must be identical to the one used in the Name parameter when you add the Jobserver to the DGC service. The value of this setting is case-sensitive. Example `-Dproxy.url=https://<your-environment-url>/reversehttp-poll/Jobserver-1`

Setting	Description
target.url	The URL of your the target system, either an on-premises Jobserver or a Tableau server. Example ◦ Jobserver: `-Dtarget.url=http://localhost:4404` ◦ Tableau server: `-Dtarget.url=https://tableau-sales.yourcompany.com`
http.proxy.host (optional)	The hostname of the HTTP proxy server for outbound connections to your Collibra Data Intelligence Cloud environment. This option is used to enable outbound traffic monitoring. Example `-Dhttp.proxy.host=proxy.yourcompany.com`
http.proxy.port (optional)	The port of the HTTP proxy server for outbound connections to your Collibra Data Intelligence Cloud environment. This option is used to enable outbound traffic monitoring. Example `-Dhttp.proxy.port=8080`

Setting	Description
username	The username of any Collibra user for basic authentication. Example `-Dusername=john.fisher`
password	The corresponding password of the Collibra user for basic authentication. Example `-Dpassword=ChangeMe` You can encrypt this password if necessary. Example `-Dpassword=enc_2:t2rklBY6699aWV0...`
keystore.path	The full path to the PKCS12 keystore. This keystore should contain the private key to sign the basic authentication header. Example `-Dkeystore.path=/opt/collibra_data/spark-jobserver/security/jobserver-1-keystore.p12`

Setting	Description

keystore.alias	The alias of the private key in the keystore. Each alias must be unique in your configuration.
	
	If you used the `name` argument during the creation of the keystore, then use the value of this `name` argument.
	
	If only 1 keystore is created, the default alias is "1".
	
	Example
	`-Dkeystore.alias=1`
	`-Dkeystore.alias=MyJobserver`

keystore.password (optional)	The password to access the keystore. If the keystore is not password-protected, don't add it to the JVM settings. You can encrypt this password if necessary. Example `-Dkeystore.password=ChangeMe`
keystore.key.password (optional)	The password to use the private key, only applicable if you secured the private key with a password. If the key is not password-protected, don't add it to the JVM settings. You can encrypt this password if necessary. Example `-Dkeystore.key.password=ChangeMe`

Setting	Description
polling.backoff	The time in milliseconds between a polling failure and a next polling attempt. We recommend to not define this parameter, it then uses the default value of 5,000 milliseconds. Example `-Dpolling.backoff=10000`
max.connections.route	The maximum number of HTTP connections per route. We recommend to not define this parameter, it then uses the default value of 20. Example `-Dmax.connections.route=30`
max.connections.total	The maximum number of all HTTP connections. We recommend to not define this parameter, it then uses the default value of 40. Example `-Dmax.connections.total=60`
idle.connection.timeout	The time in milliseconds that an idle connection is kept in the connection pool. We recommend to not define this parameter, it then uses the default value of 5,000 milliseconds. Example `-Didle.connection.timeout=3000`

Setting	Description
connection.timeout	The time in milliseconds that the reverse HTTP server waits for a response from your Collibra environment or from the value in target.url. If you don't set this parameter, the value is 60,000 milliseconds. Example `-Dconnection.timeout=30000`
connection.soTimeout	The time in milliseconds that the reverse HTTP server waits for a response from your Collibra environment or from the value in target.url on socket level. If you don't set this parameter, the value is 60,000 milliseconds. Example `-Dconnection.soTimeout=30000`
polling.timeout	The time in milliseconds that the reverse HTTP server waits for a poll request from your Collibra environment to be submitted to the target.url. If you don't set this parameter, the value is 300,000 milliseconds. Example `-Dpolling.timeout=100000`

Setting	Description
polling.period	The time in milliseconds that the reverse HTTP server waits in between poll request sessions. In other words, after having received a poll request or no request from your Collibra environment, the reverse HTTP server waits a certain amount of milliseconds before contacting the Collibra environment again. If you don't set this parameter, the value is 100 milliseconds. Example `-Dpolling.period=200`
health.check.period (optional)	The time in milliseconds that the reverse HTTP server waits between health checks of its connection with Collibra. If you don't set this parameter, the value is 5,000 milliseconds. Example `-Dhealth.check.period=10000`
health.check.timeout (optional)	The time in milliseconds that the reverse HTTP server waits for a health check response from Collibra. If you don't set this parameter, the value is 5,000 milliseconds. Example `-Dhealth.check.timeout=10000`

Note You have to use separate Jobservers for the ingestion of S3 or JDBC data sources and Tableau server data.

7. Click the green **Save all** button.
8. Click **Security configuration**.
9. Click **Edit configuration**.
10. Set the **Authentication level** to *NONE*.

> Note This means that there is a one-way outbound communication over TLS from the Jobserver to the Collibra environment, not that there is no authentication at all.

11. Click the green **Save all** button.

Add the Jobserver to the DGC service

Execute the following steps in Collibra Console of your Collibra Data Intelligence Cloud environment.

1. Open the DGC service settings for editing:
2. Go to the **Jobserver** section of the configuration.
3. Enter the required information.

Setting	Description
Name	The name of the Jobserver as it will appear when you register a data source. The name is a freely chosen name but it is recommended to only use alphanumerical characters and dashes, for example Jobserver-1. You will have to use this name as the ID of the gateway and in the address of this configuration.
Protocol	The protocol for this configuration has to be *HTTP* and not HTTPS.

Setting	Description
Address	The loopback address of the DGC service, followed by */reversehttp/<gateway-id>*. The "gateway-id" must be identical to the one used in the Name parameter of this configuration. Do not use the scheme in the address. Example localhost:4400/reversehttp/Jobserver-1
Trusted server CA certificate	The certificate in PEM format that contains the public key of the Jobserver to validate the signature of the basic authentication header. In the example to create a keystore, this is the content of the file **cert.pem**. Example -----BEGIN CERTIFICATE----- MIICqDCCAZACCQCcy3Oq51c5YzANBgkqhkiG9w0BAQsF ADAWMRQwEgYDVQQDDAtq b2JzZXJ2ZXIt... -----END CERTIFICATE-----
Client certificate	This field is not used in this configuration.
Client private key	This field is not used in this configuration. Note This field always shows dots, even if it is empty.
Table profiling data size	The approximate maximum disk size of the data in MB that will be used to profile a table. The value cannot exceed 10 000.

4. Click the green **Save all** button.

If all settings and communication paths are correctly configured, you will see a notice on the Jobserver:

```
INFO [I/O dispatcher 1] reversehttp.gateway.PollingController -
proxy -> no requests polled (204)
```

What's next?

When you have set up this communication, you may want to monitor the outbound traffic. You can do so by enabling a man-in-the-middle proxy.

Encrypt passwords for basic authentication

In the Jobserver service configuration, you have to enter an encrypted password. To encrypt the password, use the **reversehttp-gateway-standalone** utility.

Prerequisites

- You have downloaded the reversehttp-gateway-standalone-6.3.7.jar file.
- You have the password of the Collibra user that you use to connect to your Collibra Data Intelligence Cloud environment.

Steps

> Note For security reasons, we have truncated the encrypted password in the example.

1. Open a terminal or command prompt session.
2. Go to the folder that contains the downloaded JAR file.
3. Execute the following command:

```
java -jar reversehttp-gateway-standalone-6.3.7.jar encrypt
```

```
Collibra Reverse HTTP Gateway
Enter value to encrypt: <password of Collibra user>
Re-enter value to encrypt: <password of Collibra user>
Encrypted value: encrypted:k7ScuJ3...
```

> Note If the entered values in this command don't match, it will ask the values again.

What's next?

If you use an encrypted password in a configuration, use the full string of the **Encrypted value** result. This includes the prefix "encrypted:".

Monitor outbound traffic

If you set up the communication from your on-premises Jobserver to your Collibra Data Intelligence Cloud environment, you may want to monitor the outbound traffic. You can do so by setting up a man-in-the-middle proxy (MITM proxy).

1. Open Collibra Console with a user profile that has the **SUPER** role.
   » Collibra Console opens with the **Infrastructure** page.
2. In the tab pane, click the Jobserver service of a Collibra environment.
3. Click **Infrastructure Configuration**.
4. Click **JVM configuration**.
5. Click **Edit configuration**.
6. Add the following JVM settings:

Setting	Description
http.proxy.host	The hostname of the HTTP proxy server for outbound connections to your Collibra Data Intelligence Cloud environment. Example -Dhttp.proxy.host=proxy.yourcompany.com

Setting	Description
http.proxy.port	The port of the HTTP proxy server for outbound connections to your Collibra Data Intelligence Cloud environment. Example -Dhttp.proxy.port=8080

7. Add the CA certificate of this MITM proxy in the Jobserver's truststore (`$(COLLIBRA_DIR)/jre/lib/security/cacerts`).

Generate keys, certificates and keystores

For a secure communication between the Jobserver and Collibra Data Intelligence Cloud, you can use certificates. In the current configuration, certificates are used as containers for public keys and the keystore is used to store private keys and certificates.

- On the node that hosts the Jobserver service, the keystore must be in PKCS#12 format.
- On the node that hosts the Data Governance Center service, you need a certificate, in PEM format, which includes the public key.

Steps

> Note The commands used in this procedure are only examples, ask your Security officer for more information.

1. On the node on which you want to install the keystore, certificate and private key, open a terminal or command prompt session.
2. Go to or create a directory in which you want to create the keystore.
3. Create the private key and certificate:

```
openssl req -x509 -newkey rsa -keyout key.pem -out cert.pem
-days 365
```

```
Generating a 2048 bit RSA private key
.....................................+++
...................+++
writing new private key to 'key.pem'
Enter PEM pass phrase: <optional password>
Verifying - Enter PEM pass phrase: <repeat password>
-----
You are about to be asked to enter information that will be
incorporated
into your certificate request.
What you are about to enter is what is called a Distin-
guished Name or a DN.
There are quite a few fields but you can leave some blank
For some fields there will be a default value,
If you enter '.', the field will be left blank.
-----
Country Name (2 letter code) [AU]:US
State or Province Name (full name) [Some-State]:New York
Locality Name (eg, city) []:New York
Organization Name (eg, company) [Internet Widgits Pty
Ltd]:Collibra
Organizational Unit Name (eg, section) []:
Common Name (e.g. server FQDN or YOUR name) []:Francois
Lemaire
Email Address []:francois.lemaire@collibra.com
```

4. Create a PKCS#12 keystore including a private key and certificate.

```
openssl pkcs12 -export -inkey key.pem -in cert.pem -out key-
store.p12 -name <meaningful name>

Enter pass phrase for key.pem:<if password added in pre-
vious step>
Enter Export Password:
Verifying - Enter Export Password:
```

> **Important** We recommend that you provide the `name` argument with a meaningful name. You then have to use this name as the keystore alias in the JVM configuration of the Jobserver service. If you don't use the `name` argument and there's only one keystore, then the keystore alias is 1.

5. Copy the **p12** file to **%collibra_data%/spark-jobserver/security/**.

Add a Jobserver to the DGC service

To register a data source and create a data profile in Collibra Data Intelligence Cloud, you need the Jobserver service.

If you don't have a Jobserver installed and configured in your environment, the **Register data source** action will be grayed out in the global create menu of Collibra Data Intelligence Cloud.

> Tip Execute this procedure on Collibra Console of your cloud environment. In this configuration, the DGC service will send jobs to the on-premises Jobserver, however we highly recommend to revert this communication path so that the Jobserver polls the DGC service for jobs.

Steps

1. Open the DGC service settings for editing:
2. In the **Jobserver** section, click **Add**.

3. Enter the necessary information:

Setting	Description
Jobserver list	The list of registered Jobserver instances.
Name	The name of the Jobserver as it will appear when you register a data source in Data Catalog. The name is a freely chosen name but it is recommended to only use alphanumerical characters and dashes, for example Jobserver-1. You will have to use this name as the ID of the gateway and in the address of this configuration.
Protocol	The protocol that is used for the communication between the Data Governance Center service and the Jobserver service. It is recommended to use HTTPS, especially if the services are hosted in different network segments.
Address	The address (IP address, URL, hostname) of the Jobserver.
Trusted server CA certificate	The certificate of the trusted CA needed to validate the server certificate. If blank, the default truststore will be used. The default truststore is defined in the SSL configuration section of the DGC service. The CA certificate of the server party (Jobserver).
Client certificate	The client certificate offered by the DGC service to the server. If blank, you cannot select mutual authentication as the Jobserver service authentication level.
Client private key	The private key of the DGC service's certificate.

Setting	Description
Table profiling data size	The approximate maximum disk size of the data in MB that will be used to profile a table. The value cannot exceed 10,000.
Test connection timeout	This timeout is a time limit (in seconds) after which the connection test is stopped and a timeout error is shown. The default value is 60 seconds.

4. Click the green **Save all** button.

> Tip You can add as many Jobserver services as you want.

Install the Monitoring service

The Monitoring service allows you to gather metrics from Collibra Data Governance Center. The service also provides extensive monitoring and diagnostics capabilities.

For an on-premises Jobserver in combination with a Collibra Data Governance Center environment, it is optional to install an on-premises Monitoring service.

> Tip For the installation on Linux without root permissions, also read the services section.

Steps

> Note If you want to configure the init daemon on Linux systems, you have to execute an unattended installation. For more information, see also the unattended installation configuration parameters.

> Note Anti-virus and/or security software may block the installation on Windows. Make sure that these allow the installation of software and services.
> For more information, see also the Collibra University course.

1. Run the installer:
   - Linux as user with sudo rights: `sudo ./dgc-linux-5.7.12-0.sh`
     Linux as root user: `./dgc-linux-5.7.12-0.sh`
   - Linux as standard user: `./dgc-linux-5.7.12-0.sh`
   - Windows Server: double-click **setup.bat**

     > Important The path of the installer file cannot contain spaces.

     If you run the installation without Administrator rights, an error is shown.

   > Tip If you don't want to use the user interface even if it's available, add the following to the command:
   > `-- --nox11`

2. In the wizard introduction, click **Next**.
3. Enter the **Installation directory** of the Monitoring service.
   - Default location on Linux as root or user with sudo privileges: **/opt/collibra**
   - Default location on Linux as standard user: **~/collibra**
   - Default location on Windows Server: **C:\collibra**

     > Important On Windows, the target installation directory cannot contain spaces.

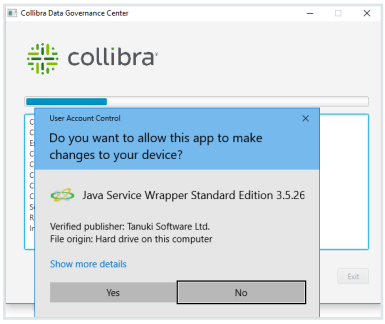
4. Click **Next**.
5. Enter the location of the **Collibra Data Directory**.
   - Default location on Linux as root or user with sudo privileges: **/opt/collibra_data**
   - Default location on Linux as standard user: **~/collibra_data**
   - Default location on Windows Server: **C:\collibra_data**

     > Important On Windows, that target data directory cannot contain spaces.

6. Click **Next**.
7. Clear all components except **Monitoring**.
8. Click **Next**.
9. Enter the port of the service. The default port is *4407*.
   If you run multiple environments on one node, this port must be unique for each environment.
10. Click **Next**.
11. Enter the Agent service settings and click **Next**.

Setting	Description
Agent port	The TCP port that is used by Collibra Console to manage the services of an environment. The default port is *4401*. If you run multiple agents on one node, this port must be unique for each agent.
Node address	The hostname of the node on which the Agent service is running. **Warning** Do not use the loopback address.

12. Click **Install**.
    » The installation of the Monitoring service starts.
13. On Windows, you may see User Account Control warnings requesting to make changes to your device.



    Click **Yes** for each of the requests, if you click **No**, the installation will fail.
14. Click **Exit**.
    » The Monitoring service is installed on your node.

What's next?

Add the Monitoring service to your environment in Collibra Console.

Upgrade the Jobserver and Collibra Console

When your Collibra Data Intelligence Cloud is upgraded, you have to update all your on-premises Jobservers and Collibra Console instances to the latest available on-premises version. The installers are released on a quarterly basis, check the compatibility list to know which installer you have to download.

This section describes how you can upgrade your on-premises Jobservers and Collibra Console.

> Tip If you already installed the on-premises Jobserver and Collibra Console with the latest available installer, there's no need to upgrade these.

Prerequisites

- You have downloaded the latest Jobserver-only installer from the Collibra Community Downloads page.

> Note
> - You must upgrade with the same user account that was used for the installation, both on Linux and Windows. If the user account is no longer active, see Upgrade an environment with another user account in the Troubleshooting section.

Steps

Linux

1. Stop the environment.
2. Stop the Collibra Agent and Collibra Console.
3. Run the installer:
   - Linux as user with sudo rights: `sudo ./dgc-linux-jobserver-only-2022.5.0-[to be defined].sh`
     Linux as root user: `./dgc-linux-jobserver-only-2022.5.0-[to be defined].sh`
   - Linux as standard user: `./dgc-linux-jobserver-only-2022.5.0-[to be defined].sh`
   - Windows Server: double-click **setupJobServerOnly.bat**
     If you run the installation without Administrator rights, an error is shown.
4. Follow the command-line wizard. If you don't enter a value, the value between brackets or the value in capital is used.

```
Verifying archive integrity...  100%   All good.
Uncompressing DGC Installer  100%
Specify the installation directory [/opt/collibra]:

/opt/collibra contains a previous installation. Do you want
to perform an update? [y/N]
y
Before you can schedule an upgrade, you need to:
- Create a backup of the entire environment.
- In Collibra Console, stop all running services on this
node.
Have you completed these steps? [yes,NO]

yes
Are you sure these are the components you want to add?
[]
The following components will be updated:
[Management Console, Jobserver] [Y/n]

2021-02-16 01:57:24.472 - SUCCESS - Check umask settings
2021-02-16 01:57:24.488 - SUCCESS - Create installation and
data directories
```

```
...

2021-02-16 01:57:46.177 - SUCCESS - Start Console
2021-02-16 01:57:46.177 - COMPLETED - Installation finished
in 22028ms.
```

5. Start the Jobserver.

Windows

1. Stop the environment.
2. Stop the Collibra Agent and Collibra Console.
3. Run the installer: Windows Server: double-click **setup.bat**The path of the installer file cannot contain spaces.
   If you run the installation without Administrator rights, an error is shown.
4. Click **Next**.
5. Select the installation directory of the old version and click **Update**.
6. Click **Yes** to confirm that you have created a backup and that all the services are stopped on the node.
   » The **Component selection** dialog box appears, indicating which services are installed on the node.
7. Click **Update**.
   » The installed services on the node are upgraded.
8. Click **Exit**.
9. Start Collibra Console.
10. Open Collibra Console with a user profile that has at least the **ADMIN** role.
    » Collibra Console opens with the **Infrastructure** page.
11. Start the Jobserver.

> Note If you encounter a communication error between the Jobserver and your environment after the upgrade, see Communication issue after upgrade from Jobserver 2021.02.

Email configuration for Collibra Data Intelligence Cloud

The Collibra cloud servers use Mandrill to reliably and securely send emails to the customer inboxes. Mandrill is powered by MailChimp, one of the leading companies in this area. For more information about the email security, see the Mandrill/MailChimp pages.

As of version 5.7, the email settings of a Collibra Data Intelligence Cloud environment are preconfigured.

By default, the From-address is no-reply@collibra.com. On request, we can update the From-address to no-reply+<environment name>@collibra.com. We do not support any other changes to the From-address.

> Example We can change the From-address to no-reply+acme-dev@collibra.com.

To request this change, contact Collibra support.

Upgrading Collibra cloud environments

As a cloud customer, your Collibra customer success management contact person will reach out to you to plan the upgrade of your Collibra Data Intelligence Cloud environment to the latest version.

You can also request an upgrade yourself via a support ticket. In your request, you then have to indicate to which version you want to upgrade and the desired date and time. We will then plan the upgrade as soon as possible.

Every upgrade causes a planned downtime. To minimize the impact, we plan your upgrade outside of your normal business hours (9 AM - 5 PM in your time zone), when feasible.

> Tip If we upgrade your Collibra Data Intelligence Cloud environment and you have an on-premises Jobserver installation, you will have to upgrade your on-premises installation if possible. The installer for on-premises installations will only be available on a quarterly basis while your Collibra Data Intelligence Cloud environment can be upgraded on a monthly basis. If your Collibra Data Intelligence Cloud environment doesn't have a corresponding on-premises installer, make sure that your on-premises Jobserver and Collibra Console are installed with the latest available installer. Check the compatibility list to know which installer you have to use.

Environment upgrades

Every environment upgrade can contain new features and bug fixes. The Collibra Data Intelligence Cloud data structure can change during an upgrade, or that existing content

will be migrated as well. This is an irreversible action, unless you restore a backup from your environment prior to the upgrade.

The proposed date and time of an upgrade is communicated to you through email, at least two weeks prior to the planned upgrade. These upgrades are done over the weekend to avoid any impact on the users.

Within these two weeks, you can request to delay this upgrade by replying to the email. You can propose a later date and time, up to two months after the initially proposed upgrade date.

> Note If you don't reply to our upgrade proposal, we assume that you approve, and we will upgrade your Collibra Data Intelligence Cloud environment at the proposed date and time.

The downtime for an upgrade depends on the size and complexity of your Collibra Data Intelligence Cloud content.

After the upgrade to 2020.11, you have to upgrade the activity history. This is a one-time only action, subsequent upgrades don't require this extra step.

> Important After an upgrade of your environment, verify if you have to upgrade your Jobserver. Check the release notes to know which Jobserver version you need. See the installation section for the upgrade steps.

Chapter 10

Troubleshooting

DGC service fails to start due to invalid configuration

If Collibra detects an invalid configuration of the Data Governance Center service, it no longer automatically replaces the invalid configuration by the default configuration and the service does not start.

This situation can happen when you edit a configuration in a backup, introducing invalid data, and then you restore that backup.

In **dgc.log**, you can find a Collibra exception "configurationParsingFailed" and "DGCConfigurationServiceImpl.readDGCConfiguration (DGCConfigurationServiceImpl.java...)":

```
Caused by: com.collibra.common.exception.CollibraException: con-
figurationParsingFailed
Message: com.fasterxml.jackson.databind.JsonMappingException:
...


    at com.collibra.dgc.configuration.service.DGCCon-
figurationServiceImpl.readDGCConfiguration(DGCCon-
figurationServiceImpl.java:362)
```

Resolution

Revert the configuration changes in your backup.

Limitations

Running Collibra Data Intelligence Cloud has a couple of limitations:

- Integrations: Connecting to your company's LDAP server is only possible if the LDAP server is accessible from the outside world (from our servers).
- Sending emails is always done through the Collibra SMTP server. For this we use mandrill.
  When receiving emails from your cloud environment, you see the following:

  > Example Received: from cloud-mail.collibra.com (cloud-mail.collibra.com. [198.2.180.134])

  This can be used to filter on your mail server (either by DNS name or IP).
- Workflow action emails requires Collibra to connect to an IMAP or POP server.
  When requested, Collibra can setup a private email address for you.
- Collibra does not support a custom domain or SSL certificate to be used for the URL of the cloud environment.
  To solve this, a custom proxy server can be used on the customer side.
- Collibra automatically applies product updates when necessary.
- Because the cloud environment are publicly accessible, guest access is not possible in the cloud environment.
- On a Collibra Data Intelligence Cloud environment, you cannot use the HTTP method PATCH.
  Use the HTTP method POST with X-HTTP-Method-Override header instead.